ISSN 2616-7433 Vol. 2, Issue 11: 51-57, DOI: 10.25236/FSST.2020.021109

An Overview of Sequence Logo Technique and Potential Application Direction

Zhanlin Dong

Tianjin University School of Pharmaceutical Science and Technology, Tianjin 300072, China
E-mail: andydong1998@tju.edu.cn

ABSTRACT. *As an important and useful technique in the field of bioinformatics, sequence logo was firstly invented by Thomas D. Schneider and R. Michael Stephens in 1990, offering an effective method to visualize the information content at each position in the sequence motif of nucleic acid or protein. In this article, the feature and principle of sequence logo are explained on a biological and statistical level. The specific operation steps to construct sequence logo are introduced. After that, in order to introduce the actual application of the technique, CodonLogo is raised as an instance that applies sequence logo to analyze codon sequence. In the end, a potential application direction aimed to SARS-CoV-2, the pathogen of COVID-19, is anticipated, covering virus genome study and important proteins (pp1ab, nsp1 to nsp16, RdRp, spike proteins) study.*

KEYWORDS: *Sequence logo technique, Biological and statistical level, Codonlogo*

1. Introduction

Sequence analysis is the main issue of bioinformatics research. As the oldest tool of bioinformatics, sequence alignment is a fundamental and important technique in sequence analysis of nucleic acid and protein, through which we obtain the degree of sequence similarity to help in the study on genes, and the prediction of structure and function of proteins. [1] In a sequence alignment, with the help of specific alignment algorithm, we input two or more sequences over the same alphabet, and the output will be an alignment of the two sequences.

The procedure of nucleic acid sequence alignment is different from protein sequence alignment. Due to 4 type of base groups (A, T, C, G), nucleic acid alignment is much more simplified. On the contrary, protein sequence alignment is more complicated because of 20 types of probable amino acids on a single position. To align the protein sequences, ungapped pairwise alignment was the primary and simplest sequence alignment method. Percentage accepted mutations (PAM) matrix and the blocks substitution matrix (BLOSUM) are useful applied on protein sequences, which provide scoring systems that score different substitutions differently, as substitution matrices. [2][3] The results in PAM or BLOSUM can be used as match scores, where high score means (commonly a positive number) better compatibility and low score (commonly a negative number) means poor compatibility or even mismatch. A sum of a terms for each pair of aligned residues, and for each gap, is defined as a scoring function. [1][2] The scoring function means a record of the relative likelihood between the aligned sequences, where identities and conservative substitutions features positive terms, while non-conservative substitutions feature negative terms. [1][2]

In order to obtain the alignment which has optimal score, it is necessary to determine whether the position ought to gap or not. Here we introduce an edit graph, where alignments between the two sequences correspond to paths between the begin and end nodes of the graph. In an edit graph, each digit from two strings respectively are aligned, and all the probable scores of the alignment results are listed (match, mismatch, and gap). [1] According to the principle of optimality, the overall optimal answer to a problem is expressed in terms of optimal answer for its sub-problems. Therefore, based on edit graph, dynamic programming is developed to calculate the overall score by combining the scores of each step. Following a traceback of the path of optimal alignment score, we can output the optimal alignment. There exist different variants of dynamic programming, the most common-used of which are Smith–Waterman algorithm and Needleman–Wunsch algorithm. [1]

The dynamic programming is an efficient algorithm to obtain the sequence alignment results. In addition, some statistic terms are introduced to analyzed the alignment results. Among which, entropy (also known as information entropy or Shannon entropy, in order to differentiate with thermodynamic entropy) is a fundamental term. The entropy of a random variable is a measurement of the uncertainty of the random variable, that is to say, a measure of the amount of information required to describe the random variable. Originally raised by Claude Shannon in 1948, the entropy of a random variable can be express by the equation:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

where $H(X)$ is the value of entropy of the random variable X, $P(x_i)$ is the probability of a value of the random variable X, and b is the base of logarithm, which is commonly 2, e, or 10. In a sequence alignment, the entropy equation is modified to adapt to the specific situation in protein sequence alignment or nucleic acid sequence alignment.

$$H(l) = -\sum_{b=a}^{l} f(b,l) \log_2 f(b,l)$$

The first equation, stated in the article by T.D. Schneider and R.M. Stephens, is used to calculate the entropy value of a position in nucleic acid sequence. In the equation, $f(b,l)$ means the frequency of a type of base b (one among A, C, G, and T) or amino acid b (one among 20 types of amino acids) at position *l*. Here the value of the base of logarithm is 2, making it in bit as the unit of entropy value. Similarly, the second equation provide a calculation of entropy value of a protein sequence, where p_a is the probability of a type of amino acid *a* at the position *l*,

To provide a measurement to how one probability distribution is different from the other one, i.e. the reference probability distribution, the term of relative entropy is introduced, as known as Kullback–Leibler divergence. The relative entropy value is calculated via the equation:

$$D = (p||q) = \sum_{b=a}^{l} f(b,l) \log_2 \left(\frac{p_a}{q_a}\right)$$

where $f(b,l)$ is the probability of base *b* or amino acid *b* in the distribution *p*, and q_a is the probability of amino acid a in reference distribution *q*. The calculation result reflects the similarity of the target sequence comparing to the reference sequence, for example, the sequence database.

Although the alignment technique above provides wide-used practical method to analyze the sequences, the results need a clear visual interpretation. The technique of sequence logo offers a helpful visualization method.

2. Results and Discussion

2.1 The Construction of Sequence Logo

The sequence logo technique was firstly developed by T.D. Schneider and R.M. Stephens in 1900. [4] It manages to visualize the information content at each position in the sequence motif, by constructing a series of logos. The logos are letters stack in several columns, and each column stand for the single position of nucleic acid or protein sequence. [4]

There are several steps in sequence logo construction procedures. Firstly, align the target sequences relative to one another, based on reference to the sequence database. In that way, a table of frequencies of each nucleic acid or amino acid at each position is constructed. After that, a series of values are calculated. Initially, we need to calculate the entropy value of each position using Equation II & III. Then, the decrease of uncertainty is determined.

$R_{sequence}(l) = \log_2 N - [H(l) + e(n)]$ [4]

$e(n) = \frac{1}{\ln 2} \times \frac{s-1}{2n}$ [4]

where *l* means the position, $\log_2 N$ is the maximum uncertainty at any given position (*N*=20 for protein and *N*=4 for amino acid), and e(n) is the correction factor (where s is 4 for nucleotides, 20 for amino acids, and *n* is the number of sequences in the alignment). The following step is to calculate the height of the logo(letter) in the single position.

$$l = f(b,l) R_{sequence}(l)$$

where $f(b,l)$ means the frequency of a type of base b or a type of amino acid b at the position *l*. The essence of these two equations is similar, and there is only difference in written forms. Finally, the nucleic acids or amino acids are stacked into columns in an increasing order of their frequencies. Vertical bars serve both as junction markers and as size markers. If the researcher needs, he or she can do further processing on the finished logos, such as sequence weighting or pseudo-count correction.

2.2 The Basic Application of Sequence Logo

As mentioned in above context, sequence logo offers a helpful visualization method. The statistical numbers are visualized by the different heights of the logos. According to the previous equations, we can raise the relationship between the heights of the logos and the frequency of the letters (standing for a nucleic acid or an amino acid).

$$l = f(b,l)\left[\sum_{b=a}^{l} f(b,l)\log_2 f(b,l) + \log_2 N - e(n)\right]$$

Here we can consider the equation as a function of $f(b,l)$, and $\log_2 N$ and $e(n)$ are constant for the alignment. Apparently, it is a monotonically increasing function in its domain. Hence, the larger the frequency of a factor, the larger the height of the corresponding letter in sequence logo. In that way, the frequency of the factors can be clearly seen. In the original paper that developed sequence logo, two examples (Fig. 1 and Fig. 2) of application on the sequence of a part of bacteriophage T7 genome and a portion of the globin make it obvious to find which type of nucleic acid or amino acid is dominant on the corresponding position. [4] The high and large size of letter indicates the large frequency of the nucleic acid or amino acid on the position, i.e. the dominant and conservative type. In conclusion, the result of sequence logo reflects a series of information, including the order of predominance and the relative frequencies of the residue at every position, as well as the amount of information (relative entropy) at every position. [4]

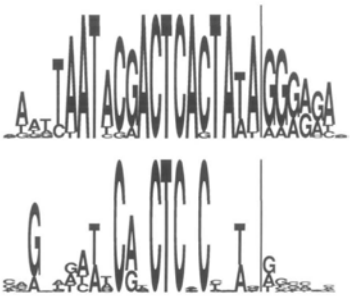

Fig.1 Logo for Sequences of T7 Rna Polymerase Binding Sites in the Bacteriophage T7 Genome. [4]

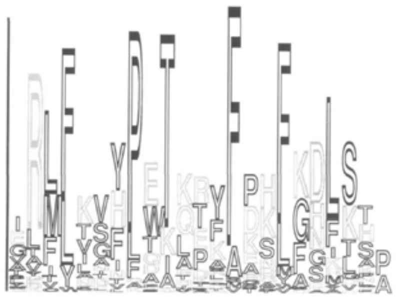

Fig.2 Logo for a Portion of the Globins. [4]

2.3 The Current Improved Application: Codonlogo

Sequence logo is a practical technique to present the data in clear visualized form. Therefore, it is helpful and appliable to all kinds of sequence alignment and analysis on protein motif, DNA, and RNA. There exists some modified application of it to adapt the specific circumstance. One of the recent modified mode is CodonLogo developed in 2012. [5]

Certain regulatory signals encoded in mRNA sequence appears as combinations of codons. Before

CodonLogo appeared, no tool was actually appliable to visualize conservative codon patterns. The problem is that it is meaningless to visualize the information on each single position of the sequence, because the codon affects in the unit of three ribosomal nucleic acid. To solve this problem, the scientists developed CodonLogo based on WebLogo3. It treats codons as inseparable entities that are parts of the alphabet of 64 types of codons. CodonLogo can discriminate and visualize conservative codon patterns from conservative nucleic acid patterns, which is indistinguishable in ordinary sequence logo. CodonLogo carries out the calculation and visualization procedures to every three positions as an integral unit. [5]

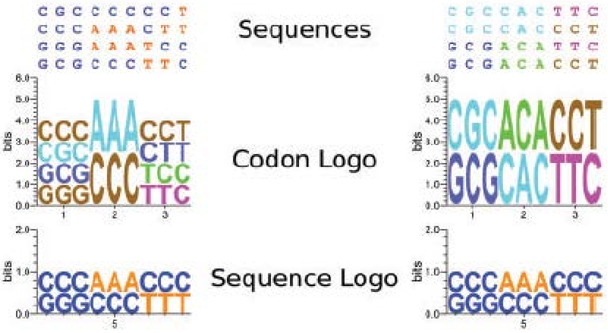

Fig.3 Different Output of Identical sequences' Alignment from Condonlogo and Ordinary Sequence Logo. [5]

From Fig.3, we can clearly find the probability dominance of specific codon, as well as corresponding amino acid refer to RNA codon table. It provides an effective method to visualize codon sequence directly. Besides, CodonLogo is able to handle the frameshift, which make it adapt to the characters of RNA, as shown in Fig.4.

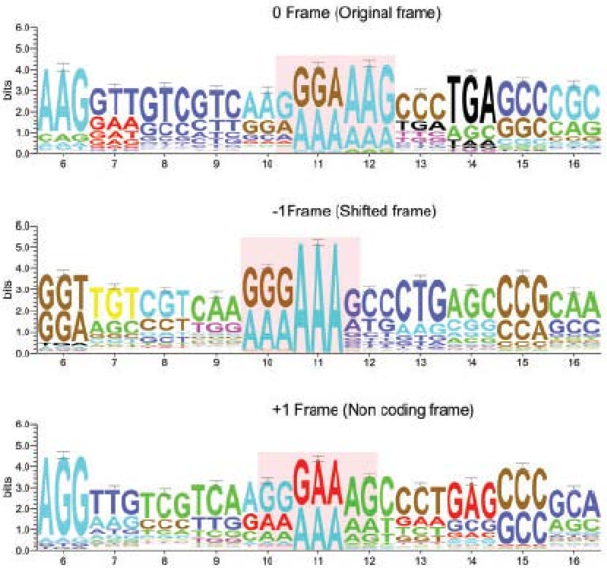

Fig.4 Codonlogo as an Alignment Result of 857 Insertion Sequences from the Is407 Family. Here the Results of Three Different Frames (Original Frame, Shifted Frame, and Non-Coding Frame) Are Presented. [5]

3. Discussion and Conclusion

Sequence logo is a method that effectively visualize the different probability of a type of nucleic acid or amino acid on the position of a sequence. It features the sequence conservation and the predominance and frequencies of the residues. The technique is wide-applied in the field of bioinformatics, to analyze and present the information within sequence. A modified mode, CodonLogo, is designed to adapt the characters of mRNA study, combining every three sequential ribosomal nucleic acid as an integral entity, a unit of codon.

4. Critical Analysis

Although sequence logo technique is applicable in bioinformatics research, it has some shortcomings and needs to be revised in the future. Firstly, the sequence logo is not a strictly quantitative presentation of the sequence information, so we cannot consider the height of letters in sequence logo as a proportional measure of conservation. Secondly, sequence logo shows less efficient ability to distinguish and visualize a series target sequence motif, which is suitable to visualize only single positions. In an analysis and presentation of a series of sequence motif in large scale, the drawback leads to uselessness. To refine it, more modified algorithm should be developed in the future, which combines a series of important sequence motif as a whole entity and visualize it.

5. Future Direction

A new direction of the application of sequence logo technique is to participate in the study of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the current official name of the pathogen of COVID-19 which is breaking out and spreading around the world, infecting millions of people. [6] As a strain of coronavirus, it has the common character of a life history with ssRNA(+), like other strains of coronavirus. As Fig.5 shows, for SARS-CoV-2, its RNA positive strand is released from the protein envelop after entering the host cell, and involve in two times RNA replication to synthesize new RNA positive strand for next generation, with the help of RNA-dependent RNA polymerase (RdRp). In another pathway, the parental RNA positive strand directly participate transformation as mRNA to synthesize protein capsid of next generation. The newly synthesized RNA positive strand and protein capsid aggregate to new virus as next generation. [7][8]

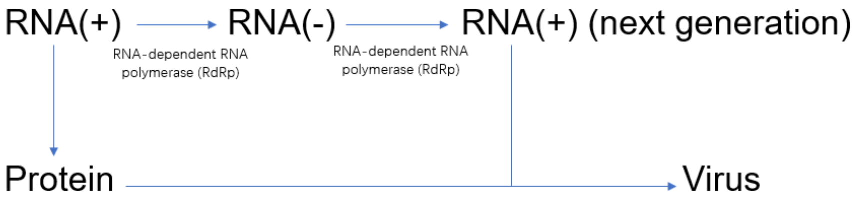
RNA(+)
RNA(-)
RNA(+) (next generation)
RNA-dependent RNA polymerase (RdRp)
RNA-dependent RNA polymerase (RdRp)
Protein
Virus

Fig.5 A Brief Schematic Diagram of the Life History of Ssrna(+) Virus.

The genome of coronavirus can be considered as a messenger RNA molecule, and two-thirds of it is translated into two large overlapping polyproteins (pp1a and pp1ab). The polyproteins contain proteases (PLpro and 3CLpro), which cleave the polyproteins themselves at specific sites and produce 16 nonstructural proteins (nsp1 to nsp16) from the cleaved pp1ab. Product proteins include a series of replication proteins, such as RNA-dependent RNA polymerase (RdRp), RNA helicase, and exoribonuclease. The nonstructural replication proteins are combined to form a multi-protein replicase-transcriptase complex (RTC). [8][9] RdRp directly mediates the transcription and replication of the viral genome. The other nonstructural proteins in the complex revolve in the process of replication and transcription. [8][9]

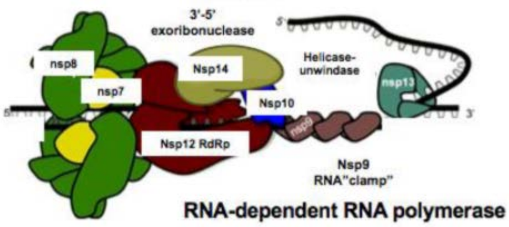

Fig.6 A Model Figure of the Replicase-Transcriptase Complex in Coronavirus. [9]

Another important structure of coronavirus is spike protein. The main structure of the viral envelope of coronavirus is a lipid bilayer, attached with the membrane (M), envelope (E) and spike (S) proteins. The spike proteins form bulbous projections on the surface of virus envelope. Their interaction with the specific receptor of complement host cell is crucial to determine the tissue tropism and infectivity. [10]

It is important to find the conservative sequence of genome and protein, in order to develop drugs or vaccines affecting on specific virus protein (polyproteins pp1a and pp1ab, proteases PLpro and 3CLpro, RdRp, nonstructural proteins, spike protein, etc.) or RNA segment. In the ongoing study of SARS-CoV-2 genome and proteins, sequence logo is sure to play an important part in sequence analysis and presentation via clear visualization.

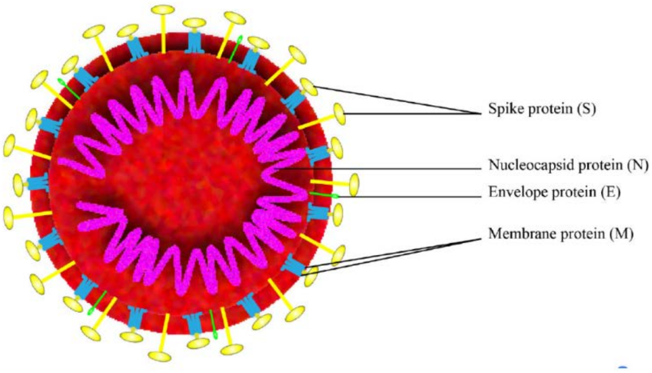

Fig.7 A Schematic Diagram of the Distribution of Different Types Protein of Coronavirus. [10]

6. Conclusion

Since developed in 1990, the technique sequence logo provides a helpful and efficient method to visualize the information within nucleic acid sequence or amino acid sequence. The procedure to construct sequence logo contains a series of calculation, featuring the order of predominance and the relative frequencies of every residue at every position, which reflects the degree of conservation. It is commonly used in sequence analysis and presentation in the field of bioinformatics. A recent modified mode called CodonLogo combines every three ribosomal nucleic acids as an entity of codon in mRNA study, making it more convenient and efficient. However, sequence logo is not a strictly quantitative visualization technique. Furthermore, it is still inefficient on the analysis of the sequence conservation in a large scale. Recently, a potential direction of sequence logo application is to take part in the study of the genome and important proteins (replicase-transcriptase complex spike protein, etc.) of SARS-CoV-2.

References

[1] Lund O., Nielsen M., Lundegaard C., Kesmir C., Brunak S (2005). Immunological Bioinformatics, The MIT Press, 2005.
[2] Dayhoff M.O., Schwartz R., Orcutt, B.C(1978).Atlas of Protein Sequence and Structure, National Biomedical Research Foundation.
[3] Henikoff S., Henikoff J.G(1992).Amino Acid Substitution Matrices from Protein Blocks, Proc Natl Acad Sci USA, vol.89, no.22, pp.10915-10919.
[4] Schneider T.D., Stephens R.M (1990). Sequence Logos: A New Way to Display Consensus Sequences, Nucleic Acids Research, vol.18, no.20, pp. 6097-6110.
[5] Sharma V., Murphy D.P., Provan G., Baranov P.V (2012). CodonLogo: A Sequence Logo-based Viewer for Codon Patterns, Bioinformatics, vol.28, no.14, pp.1935-1936.
[6] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nature Microbiology, no.5, pp.536-544.
[7] Nagy P.D., Pogany J (2012).The Dependence of Viral RNA Replication on Co-opted Host Factors, Nat Rev Microbiol, vol.10, no.2, pp.137-149.
[8] Fehr A.R., Perlman S (2015).Coronaviruses: Methods and Protocols, Humana Press.
[9] Huang J.S., Song W.L, Huang H (2020). Pharmacological Therapeutics Targeting RNA-Dependent RNA Polymerase, Proteinase and Spike Protein: From Mechanistic Studies to Clinical Trials for COVID-19, Journal of Clinical Medicine, vol.9, no.4, pp.1131.
[10] Zhang J.Y., Zeng H., Gu J., Li H.B (2020). Progress and Prospects on Vaccine Development against SARS-CoV-2, Vaccines, vol.8, no.2, pp.153-154.